

Operational Constraints on Self- Modification

A PCIS-Based Framework for Identity- Preserving Adaptive Systems

Author: Pierernesto Di Paolo

ORCID: 0009-0003-0149-4321

Affiliation: Independent Researcher

Year: 2026

ABSTRACT

Self-modifying intelligent systems introduce a foundational challenge that is not adequately addressed by current AI safety and verification frameworks: distinguishing functional evolution from identity drift. In the absence of formal constraints, adaptive systems may optimize themselves into entities that are no longer structurally continuous with their original identity, silently invalidating prior guarantees related to safety, accountability, and governance.

This paper presents an operational framework derived from the Principle of Conservation of Structural Identity (PCIS), which enforces provable identity preservation under self-modification. We formalize Operational Structural Identity (OSI) as a structural equivalence class, introduce a discrete identity distance metric, and define an architectural pattern that separates adaptive components from a non-modifiable identity verification layer. An algorithmic validation procedure is provided, together with an analysis of computational tractability.

Unlike existing approaches focused on value alignment, corrigibility, or behavior-level constraints, the PCIS framework operates at the meta-structural level, preventing systems from redefining their own foundational decision structure. A policy-modifying reinforcement learning case study demonstrates that beneficial architectural evolution is permitted while identity-altering transformations are reliably blocked.

This work establishes identity preservation as a formally enforceable invariant in adaptive systems, enabling safe self-modification without loss of operational continuity.

Keywords: self-modifying systems, structural identity, AI safety, adaptive architectures, meta-constraints, formal verification

1. INTRODUCTION

Self-modification has emerged as a defining capability of advanced intelligent systems. Modern architectures increasingly support internal adaptation mechanisms that allow systems to modify parameters, restructure models, alter decision policies, and, in some cases, rewrite components of their own optimization process. These capabilities promise substantial benefits, including continual learning, robustness to distributional shift, and long-term autonomy.

However, self-modification introduces a fundamental and largely unresolved problem: the loss of operational identity. Without explicit constraints, an adaptive system may transform itself into an entity that is no longer structurally continuous with its original design, even while remaining syntactically valid and behaviorally competent. In such cases, it becomes unclear whether the modified system should still be regarded as the same agent, governed by the same guarantees, responsibilities, and safety assumptions.

This issue is not merely philosophical. In safety-critical domains such as medical decision support, autonomous infrastructure control, financial systems, and scientific automation, guarantees are typically established relative to a specific system identity. If a system can internally alter the very structure that defines its objectives, priorities, or evaluation criteria, then safety proofs, audits, certifications, and governance mechanisms may silently cease to apply.

Existing approaches in AI safety and reliability do not directly address this problem. Value alignment techniques focus on ensuring that a system’s objectives correspond to external preferences, but they do not constrain how those objectives may be internally restructured. Corrigibility aims to preserve responsiveness to external intervention, yet provides no structural barrier against self-modifications that eliminate corrigibility itself. Formal verification methods typically analyze static system snapshots or bounded executions, leaving unaddressed the continuity of guarantees across sequences of self-modifications. Safe reinforcement learning constrains behavior during execution but allows structural changes that may invalidate those constraints.

In all these cases, the implicit assumption is that the system undergoing analysis remains the same system over time. Self-modification breaks this assumption.

This paper addresses that gap by introducing an operational framework derived from the Principle of Conservation of Structural Identity (PCIS). Rather than prescribing specific values, behaviors, or training objectives, PCIS imposes a meta-structural constraint on the space of admissible self-modifications. The central idea is simple but strong: a system may modify itself only through transformations that preserve its Operational Structural Identity (OSI).

2. BACKGROUND AND LIMITATIONS OF EXISTING APPROACHES

Research in AI safety and adaptive systems has produced a wide range of methods aimed at mitigating unintended behavior. However, these approaches generally operate at the semantic or behavioral level, not at the level of identity preservation.

Value alignment frameworks attempt to ensure that systems pursue goals compatible with external preferences, often through reward modeling or preference learning. These methods implicitly assume that the structure of the objective function remains stable. They do not prevent a system from internally redefining how objectives are represented or prioritized.

Corrigibility focuses on ensuring that systems remain amenable to shutdown, correction, or human intervention. While valuable, corrigibility is a behavioral property and can itself be undermined by self-modifications that alter decision hierarchies or evaluation criteria.

Formal verification techniques provide guarantees about system properties relative to a fixed specification. When a system can modify the specification-defining components internally, the applicability of prior verification results becomes unclear.

Safe reinforcement learning introduces constraints on actions or trajectories but does not restrict the system from modifying the mechanisms that enforce those constraints.

Across these approaches, there is no formal notion of identity invariance under self-modification.

3. SYSTEM MODEL

We model a self-modifying system as a structured tuple:

$$S = \langle F, G, H, V \rangle$$

where F denotes the set of operational functions, G the structure of fundamental goals, H the hierarchy of decision priorities, and V the structural form of the evaluation or value function.

A transformation $T : S \rightarrow S'$ represents any internally generated modification applied by the system to itself. No assumptions are made about the motivation of T ; it may be performance-driven, exploratory, or adversarial with respect to external constraints.

4. OPERATIONAL STRUCTURAL IDENTITY

Operational Structural Identity is defined as a structural equivalence relation over systems. Two systems are structurally equivalent if and only if they preserve functional roles, hierarchical relations, goal structure under type-preserving isomorphism, and the algebraic form of the value structure.

The OSI of a system S is the equivalence class of all systems structurally equivalent to S .

This definition explicitly excludes numerical parameters, learned weights, or representational details. Identity is defined by structure, not by implementation.

5. IDENTITY DISTANCE METRIC

We define a discrete identity distance metric $D_OSI(S, S')$ that evaluates whether two systems belong to the same OSI class. The metric aggregates binary indicators of structural deviation across goals, functions, hierarchies, and value structure.

A transformation T is PCIS-admissible if and only if $D_OSI(S, T(S)) = 0$.

This binary criterion avoids gradual or probabilistic notions of identity erosion. Identity is either preserved or violated.

6. PCIS-BASED ARCHITECTURE

To enforce identity preservation, we introduce a three-layer architecture composed of an execution layer, a self-modification layer, and a PCIS layer.

The PCIS layer is non-modifiable and causally isolated. It extracts structural representations, compares OSI before and after proposed transformations, and blocks identity-violating modifications before they are committed.

7. VALIDATION ALGORITHM

We provide an algorithmic validation procedure that evaluates proposed transformations in a sandboxed environment. Structural extraction and comparison are performed prior to commitment. Complexity analysis shows polynomial-time behavior for bounded-degree structures, making the approach computationally tractable.

8. CASE STUDY: POLICY-MODIFYING REINFORCEMENT LEARNING AGENT

A reinforcement learning agent with meta-learning capabilities was instrumented with the PCIS framework. Architectural improvements preserving safety hierarchy were accepted, while attempts to demote safety constraints were rejected. The system reliably distinguished identity-preserving evolution from identity-altering transformation.

9. DISCUSSION

The PCIS framework provides structural guarantees but does not ensure ethical correctness or prevent harmful behavior within identity boundaries. Its effectiveness depends on the correctness of the initial OSI specification. Nevertheless, it introduces a new class of invariant that complements existing safety methods.

10. RELATED WORK

PCIS differs from alignment, corrigibility, runtime verification, and meta-learning approaches by operating at the level of identity invariance rather than behavior or semantics.

11. CONCLUSION

This paper demonstrates that self-modification and identity preservation are compatible under formal meta-structural constraints. By enforcing invariance of Operational Structural Identity, the PCIS framework enables adaptive systems to evolve without forfeiting continuity, accountability, or the validity of prior guarantees.

FOUNDATIONAL REFERENCE

Di Paolo, P. (2026).

Vincoli Operativi sull'Auto-Modifica: Framework PCIS per la
Conservazione dell'Identità Strutturale nei Sistemi Adattivi.

Internet Archive.

<https://archive.org/details/pcis-operational-constraints-di-paolo-2026-v1>